

# **EXHIBIT J**

UNITED STATES DISTRICT COURT  
NORTHERN DISTRICT OF CALIFORNIA  
SAN FRANCISCO DIVISION

RICHARD KADREY, et al.,

*Individual and Representative Plaintiffs,*

v.

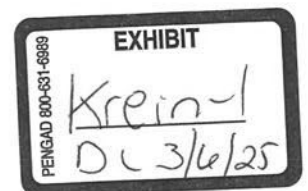
META PLATFORMS, INC.,

*Defendant.*

Case No. 3:23-cv-03417-VC

**OPENING EXPERT REPORT OF  
DR. JONATHAN L. KREIN**

**HIGHLY CONFIDENTIAL – ATTORNEYS’ EYES ONLY AND SOURCE CODE**



## HIGHLY CONFIDENTIAL – ATTORNEYS’ EYES ONLY AND SOURCE CODE

A. A portion of -- in all of these there was a portion of, because it's not the same -- it's not the complete set of data after it goes through the cleaning process. And so the Books3 was used in pretraining in LLaMA 2.

Q. And what about LLaMA 3?

A. Yes, Books3 was used in the training of LLaMA 3.

Q. And what about LLaMA 4?

A. Yes, Books3 is in -- that is just entering pretraining, and the last data mix that I am aware of, that was included.

## 8.2. Meta Has Directly Downloaded At Least One Copy of Books3

71. In my review of Meta's source code and documentation, I have found evidence that Meta directly downloaded at least one copy of Books3. In this section, I cite source code and supporting documentation that establish direct downloading of Books3 by Meta.

72. Meta source code has been identified that is used to directly download The Pile,<sup>28</sup> which includes Books3.<sup>29</sup> In the file `pile_download.py`, a source code comment states “download shard zst file from The pile.”<sup>30</sup> The downloading of specific files from The Pile is performed by the function `download_file()`.<sup>31</sup> Notice as well that the path for this code file includes the folders “.../projects/pile/...”<sup>32</sup>

73. A spreadsheet provided by Meta shows the downloading and/or use of data from shadow libraries including Books3 (in this case, as part of b3g) for developing Meta's Llama 1,

---

<sup>28</sup> fair\_data [REDACTED] (META-KADREY-SC-000434).

<sup>29</sup> See Appendix A.

<sup>30</sup> fair\_data [REDACTED] (META-KADREY-SC-000434) at line 72.

<sup>31</sup> fair\_data [REDACTED] (META-KADREY-SC-000434) at lines 26-34.

<sup>32</sup> E.g., fair\_data [REDACTED] See META-KADREY-SC-000434.

## HIGHLY CONFIDENTIAL – ATTORNEYS’ EYES ONLY AND SOURCE CODE

Llama 2, and Llama 3 models.<sup>33</sup> According to the document’s metadata, the spreadsheet file is named “Copy of LLM Pretraining data tracker.” The first tab is labeled, “Oct-Nov 2023 Data Planning.” The spreadsheet includes 29 tabs that articulate planning, status, and results pertaining to and/or stemming from the direct downloading of Books3 (among other datasets) going back to Llama 1. The spreadsheet also includes an initial outlay for Llama 4.

74. As one example from the spreadsheet showing that Meta downloaded Books3 at least once, the tab labeled “Llama 3 data” includes an entry in cell A12 that reads, “Data for llama 3.” In the area associated with that entry, under a column header labeled “Dataset” (see cell G13), the data for Llama 3 is specified to include, among others, “b3g (llama 1/2)” (see cell G15).<sup>34</sup> Thus, as for numerous other references contained in the spreadsheet, this one shows b3g, including at least one copy of Books3, was downloaded by Meta and used in the development of Llama 1, Llama 2, and Llama 3.

### 8.3. Meta Has Created Numerous Internal Copies of Books3

75. Post download, prolific copying of Books3 occurs at various stages throughout the machine learning pipeline including for purposes of data curation, model training, fine tuning, and model evaluation. This section discusses Meta’s source code that demonstrates copying during these processes.

76. Training data, whether from shadow libraries, web crawls, or other sources, comes in many formats, including HTML, PDF, and EPUB, to name a few. This data goes through a curation process and is ultimately converted into “JSON Lines” files (.jsonl) which contain one

---

<sup>33</sup> Meta\_Kadrey\_00042928.

<sup>34</sup> See Meta\_Kadrey\_00042928.

HIGHLY CONFIDENTIAL – ATTORNEYS’ EYES ONLY AND SOURCE CODE

**8.6. Meta Has Removed Copyright Information from Copies of Books3**

95. Meta admits to stripping copyright notices from copyrighted books during the development of various Llama models. Nikolay Bashlykov, for example, testified:<sup>66</sup>

Q. Yes. So what does the fair\_data source code repository contain?

A. So, in general terms, it contains processing scripts, steps for various datasets.

Q. So would the script that you created to strip the copyright information from books be contained in fair\_data?

A. Correct.

96. My review of Meta’s source code shows that Meta created scripts to remove copyright information—such as copyright notices, ISBNs, author information, and owner information—including from Books3, as an early step in the data curation process. My review further finds, owing to the numerous instances of such scripts, that Meta likely created and preserved dozens, possibly hundreds of copies of copyrighted works having their copyright notices removed, including potentially many from Books3.

97. For example, the script `pile_clean_v0.py` removes copyright information from The Pile dataset (which includes Books3).<sup>67</sup> The `main()` function sets up “cleaner” functions, each for their respective dataset (*i.e.*, “ArXiv,” “Books3,” “Github,” *etc.*).<sup>68</sup> The “Books3” dataset is assigned to the “cleaner” function `fixup_Book()`.<sup>69</sup> When the “Books3” “cleaner” is executed,

---

<sup>66</sup> Bashlykov Dep. Tr. (December 6, 2024) at 142:5-12.

<sup>67</sup> See fair\_data [REDACTED] (META-KADREY-SC-000031).

<sup>68</sup> fair\_data [REDACTED] (META-KADREY-SC-000031) at lines 387-410.

<sup>69</sup> fair\_data [REDACTED] (META-KADREY-SC-000031) at lines 390, 429.



## HIGHLY CONFIDENTIAL – ATTORNEYS’ EYES ONLY AND SOURCE CODE

fixup\_Book() is accordingly invoked.<sup>70</sup> In turn, fixup\_Book() calls the raw\_bookscorpus() function,<sup>71</sup> which performs the removal of copyright information—including any “Published by,” “Copyright,” “Edited by,” “Smashwords,” “All rights reserved,” and “ISBN” information.<sup>72</sup>

98. As another example, the script **pile\_clean\_v1.py** removes copyright information from The Pile dataset (which includes Books3).<sup>73</sup> Similar to the “v0” script, **pile\_clean\_v1.py**’s main() function sets up “cleaner” functions by assigning each to a dataset (*i.e.*, “ArXiv,” “Books3,” “Github,” *etc.*).<sup>74</sup> The “Books3” dataset is assigned to the “cleaner” function book3().<sup>75</sup> When the “Books3” “cleaner” is executed, the book3() function is accordingly invoked.<sup>76</sup> In turn, book3() calls the generic\_clean\_book() function,<sup>77</sup> which is defined in the script file **text\_cleaner.py**.<sup>78</sup>

---

<sup>70</sup> fair\_data [REDACTED] (META-KADREY-SC-000031) at lines 436, 359.

<sup>71</sup> fair\_data [REDACTED] (META-KADREY-SC-000031) at line 364.

<sup>72</sup> fair\_data [REDACTED] (META-KADREY-SC-000031) at lines 21-32.

<sup>73</sup> See fair\_data [REDACTED] (META-KADREY-SC-000625).

<sup>74</sup> fair\_data [REDACTED] (META-KADREY-SC-000625) at lines 310-333.

<sup>75</sup> fair\_data [REDACTED] (META-KADREY-SC-000625) at lines 313, 352.

<sup>76</sup> fair\_data [REDACTED] (META-KADREY-SC-000625) at lines 359, 63.

<sup>77</sup> fair\_data [REDACTED] (META-KADREY-SC-000625) at line 65.

<sup>78</sup> fair\_data [REDACTED] (META-KADREY-SC-000625) at line 14; fair\_data [REDACTED] (META-KADREY-SC-000045) at lines 337-416.

HIGHLY CONFIDENTIAL – ATTORNEYS’ EYES ONLY AND SOURCE CODE

99. The `generic_clean_book()` function defines a list of keywords for use in identifying material to be removed, including:<sup>79</sup>

- Published by
- Produced by
- Copyright
- Cover copyright
- Cover designed by
- Edited by
- Prologue
- PREFACE
- Epilogue
- Acknowledgements
- About the Author
- Smashwords Edition
- All rights reserved
- Title Page
- Table of Contents
- \_Introduction\_
- \_Cover\_
- \_Title Page\_
- \_Copyright\_

Subsequently, the `generic_clean_book()` function calls the `rm_line_start_from_words()` function, passing the keywords.<sup>80</sup> Source code comments describe the `rm_line_start_from_`

---

<sup>79</sup> fair\_data [REDACTED] (META-KADREY-SC-000045) at lines 370-390.

<sup>80</sup> fair\_data [REDACTED] (META-KADREY-SC-000045) at lines 395-396, 116-133.

## HIGHLY CONFIDENTIAL – ATTORNEYS’ EYES ONLY AND SOURCE CODE

words() function as “Remove lines starts with any of entities in given word\_list.”<sup>81</sup>

The rm\_line\_start\_from\_words() function uses the keywords list to identify lines for removal.<sup>82</sup>

100. As another example, the script **BooksHeaderCleaner.scala** removes copyright information from the Books3 dataset.<sup>83</sup> The main() function calls the tryStripHeader() function,<sup>84</sup> which creates a list of keywords, including:<sup>85</sup>

- all rights reserved
- copyright
- ©
- already a subscriber
- thank you for downloading
- mcgraw-hill education
- about the author
- by the same author
- simon & schuster
- ®
- tel:
- penguin books
- harpercollins

---

<sup>81</sup> fair\_data [REDACTED] (META-KADREY-SC-000045) at line 117.

<sup>82</sup> fair\_data [REDACTED] (META-KADREY-SC-000045) at lines 128-132.

<sup>83</sup> See fair\_data\_2024-08-21 [REDACTED] (META-KADREY-SC-000037) incl. at line 160.

<sup>84</sup> fair\_data [REDACTED] (META-KADREY-SC-000037) at lines 145, 197.

<sup>85</sup> fair\_data [REDACTED] (META-KADREY-SC-000037) at lines 75, 115-117.



## HIGHLY CONFIDENTIAL – ATTORNEYS’ EYES ONLY AND SOURCE CODE

- @
- penguin group
- www.penguin
- penguingroup.com

These keywords—along with a few additional search terms defined piecemeal (*i.e.*, “isbn,” “ebook,” “e-book,” “notebook,” “note-book,” “facebook,” “casebook,” “www.,” and “introduction”)—are used to identify and filter lines.<sup>86</sup>

101. As yet another example, the script **BooksFooterCleaner.scala** removes copyright information from the Books3 dataset.<sup>87</sup> The `main()` function calls the `tryStripFooter()` function,<sup>88</sup> which creates a couple lists of keywords, collectively encompassing:<sup>89</sup>

- acknowledgments
- acknowledgements
- @
- author’s note
- index
- glossary
- bibliography
- contributors
- other works
- further reading

---

<sup>86</sup> fair\_data [REDACTED]  
(META-KADREY-SC-000037) at lines 119-126.

<sup>87</sup> See fair\_data\_2024-08-21 [REDACTED] (META-KADREY-SC-000040) incl. at line 189.

<sup>88</sup> fair\_data [REDACTED]  
(META-KADREY-SC-000040) at lines 178, 245.

<sup>89</sup> fair\_data [REDACTED]  
(META-KADREY-SC-000040) at lines 11, 24-28.

HIGHLY CONFIDENTIAL – ATTORNEYS’ EYES ONLY AND SOURCE CODE

- leave a review
- notes
- the end
- resources
- works cited
- random house
- all rights reserved
- copyright
- ©
- already a subscriber
- thank you for downloading
- mcgraw-hill education
- about the author
- note on the author
- by the same author
- simon & schuster
- ®
- tel:
- penguin books
- harpercollins
- thank you for purchasing

These keywords—along with a few additional search terms defined piecemeal (*i.e.*, “isbn,” “www.,” and “copyright”)—are used to identify and filter lines.<sup>90</sup>

---

<sup>90</sup> fair\_data [REDACTED]  
(META-KADREY-SC-000040) at lines 30-40.

## HIGHLY CONFIDENTIAL – ATTORNEYS’ EYES ONLY AND SOURCE CODE

the storage locations for copies of datasets refer to workspaces allocated to specific Meta engineers.

109. The document entitled “*LibGen dataset: 650B\* clean & deduped tokens*” also shows the downloading and use of data from shadow libraries including LibGen for developing Meta’s Llama models,<sup>103</sup> beginning at least with Llama 3 (but further showing use of LibGen in conjunction with Llama 2 as well<sup>104</sup>).

110. The front-page description of this document states in part:<sup>105</sup>

Library Genesis, or LibGen, is a search engine and digital library that provides free access to a vast collection of books, articles, and other scholarly materials. ... The platform offers PDF and EPUB (ZIP archive containing a collection of HTML, CSS, ...) versions of books and articles, often sourced from copyrighted materials without the permission of the copyright holders.

111. This document also includes a table that identifies material having been downloaded from LibGen.<sup>106</sup> In fact, column 3 of this table is labeled, “Downloaded (doc num / %).” The count of downloaded works in this column totals more than 3.5 million.

112. Later in the same document, a note by “Nikolay”<sup>107</sup> (dated “06.06.2023”) reads:<sup>108</sup>

Done with the EN Scitech/Fiction part. Now finishing the non-EN Scitech/Fiction and ALL Scimag.

Sci-tech (non-en): Downloaded 130k (99%) of non-English epub/mobi Sci-tech books and 586k non-English epub/mobi Fiction books.

---

<sup>103</sup> Meta\_Kadrey\_00065244.

<sup>104</sup> See, e.g., Meta\_Kadrey\_00065244 at 257-259.

<sup>105</sup> Meta\_Kadrey\_00065244 at 244.

<sup>106</sup> Meta\_Kadrey\_00065244 at 246-247.

<sup>107</sup> I understand this to be Nikolay Bashlykov.

<sup>108</sup> Meta\_Kadrey\_00065244 at 274.

HIGHLY CONFIDENTIAL – ATTORNEYS’ EYES ONLY AND SOURCE CODE

113. And shortly below that, another note by “Nikolay” (dated “05.06.2023”) states: “Sci-mag is 70% downloaded ... Started loading multi-lang Scitech & Fiction.”<sup>109</sup>

114. The document is also filled with download metrics specific to the downloading of LibGen and parts of LibGen.<sup>110</sup>

115. As just one example from the document showing that Meta *directly* downloaded LibGen, “Nikolay” states on “09.05.2023”:<sup>111</sup>

Decided to go with the *direct file upload without using torrents* for the following reasons:

- *using torrents would entail “seeding” the files - i.e. sharing the content outside, this could be legally not OK*
- with the direct file download we can pre-filter the needed format and language of the files - i.e. downloading only EN, PDF and EPUB initially
- *the downside is that this way it is slower and need more engineering* to bypass IP throttling and download retries
- we can reload specific MD5 file names, that were corrupted or missing from the initial download (based on Lukas's observations there are 30% of corrupted files in the initial Libgen download)

116. These instances cited within this document are just a few representative samples of many such references to the downloading and/or use of LibGen by Meta found within Meta’s source code and documents.

---

<sup>109</sup> Meta\_Kadrey\_00065244 at 275.

<sup>110</sup> See, e.g., Meta\_Kadrey\_00065244 at 246-247, 274-275, 276-277, 278, 279, 280, 284, 285, etc.

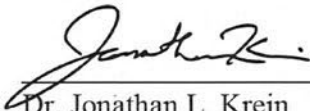
<sup>111</sup> See, e.g., Meta\_Kadrey\_00065244 at 304, emphasis added.

HIGHLY CONFIDENTIAL – ATTORNEYS’ EYES ONLY AND SOURCE CODE

I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct.

Executed on January 10, 2025

Respectfully submitted,

  
\_\_\_\_\_  
Dr. Jonathan L. Krein



UNITED STATES DISTRICT COURT  
NORTHERN DISTRICT OF CALIFORNIA  
SAN FRANCISCO DIVISION

RICHARD KADREY, AN : CASE NO.  
INDIVIDUAL; SARAH SILVERMAN, : 3:23-cv-03417-VC  
AN INDIVIDUAL; CHRISTOPHER :  
GOLDEN, AN INDIVIDUAL, :  
PLAINTIFFS :  
VS. :  
META PLATFORMS, INC., A :  
DELAWARE CORPORATION, :  
DEFENDANT :

\_\_\_\_\_ :  
HIGHLY CONFIDENTIAL - ATTORNEYS' EYES ONLY

VIDEOTAPED DEPOSITION OF JONATHAN KREIN, PH.D.  
SAN FRANCISCO, CALIFORNIA  
THURSDAY, MARCH 6, 2025

REPORTED BY:  
DEBBIE LEONARD, CSR, RDR, CRR  
CSR NO. 14350  
JOB No. 7189213

PAGES 1 to 113

Page 1

1 that you can't give your best and most accurate testimony  
2 here today?

3 A No.

4 Q And are you taking any medication that could  
5 affect the content of the testimony you're giving here  
6 today?

7 A No.

8 Q Okay. Thank you, sir.

9 And, sir, you've been retained as an expert  
10 witness on behalf of the plaintiffs in this case,  
11 correct?

12 A Yes.

13 Q Approximately when were you first retained?

14 A I do not recall. It would have been many months  
15 ago. I don't recall.

16 (Krein Exhibit 1 marked for identification.)

17 BY MR. WEINSTEIN:

18 Q And I have handed you -- or the court reporter  
19 has put in front of you Exhibit 1, which is a copy of  
20 your expert report -- opening expert report served in  
21 this case.

22 You have also served declarations previously in  
23 this case, correct?

24 A I have, yes.

25 Q With respect to expert reports, is Exhibit 1 the

1 only expert report you've served?

2 A Yes.

3 Q Okay. And in preparation for your deposition  
4 here today, did you have an opportunity to go over your  
5 report again?

6 A I did re-read it.

7 Q Okay. Thank you.

8 And I don't want to get into communications with  
9 counsel, pursuant to Rule 26, but generally speaking,  
10 what did you do to prepare for your deposition here  
11 today?

12 MR. YOUNG: And I just caution the witness, per  
13 counsel's admonition, to just be cognizant not to reveal  
14 any conversation you might have had with counsel.

15 THE WITNESS: I reviewed my report, reviewed  
16 another document or two, and spoke with counsel.

17 BY MR. WEINSTEIN:

18 Q Okay. The other document or two, can you  
19 identify what those were?

20 MR. YOUNG: I'm going to object to the fact that  
21 that calls for attorney mental impressions.

22 MR. WEINSTEIN: And I can lay a foundation for  
23 that.

24 BY MR. WEINSTEIN:

25 Q Were these documents that were furnished to you

C E R T I F I C A T E

I, Debbie Leonard, Certified Shorthand Reporter  
No. 14350 for the State of California, do hereby  
certify:

That the foregoing deposition was taken before me  
at the time and place therein set forth, at which time  
the witness was put under oath by me; that the testimony  
of the witness and all objections made at the time of the  
examination were recorded stenographically by me, were  
thereafter transcribed by me by means of computer; and  
that the foregoing is a true record of same.

I further certify that I am neither counsel for  
nor related to any party to said action, nor in any way  
interested in the outcome thereof.

IN WITNESS WHEREOF, I have subscribed my name  
this 13th day of March, 2025.



Debbie Leonard, CSR, RDR, CRR

CSR NO. 14350